20.9.2017

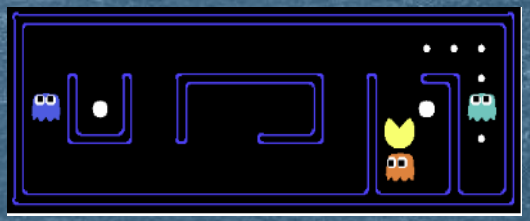
Nemanja Subotić
suboticnemanja93@gmail.com

Reinforcement learning kroz Pacman igricu

Seminarski rad iz predmeta Naučno izračunavanje
Matematički fakultet, Univerzitet u Beogradu

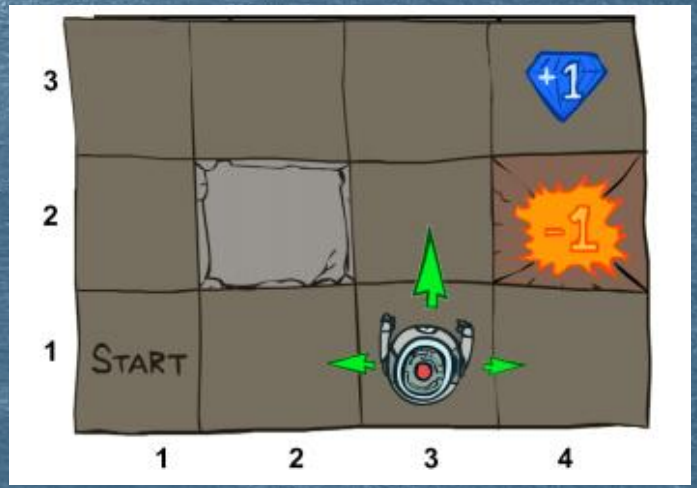
O projektu

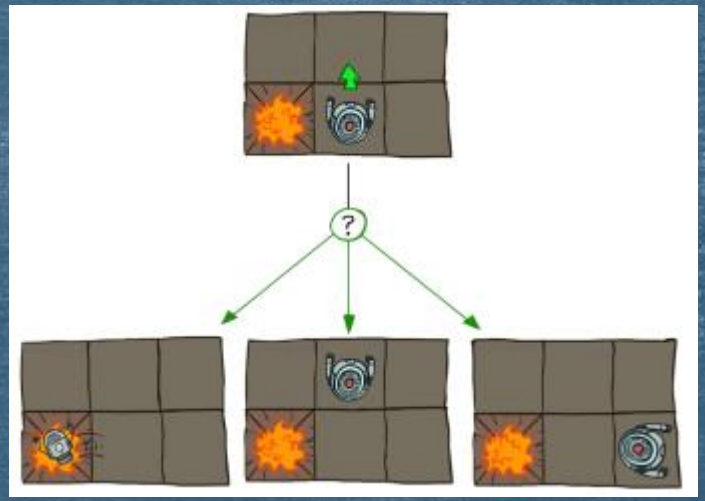
- Markovljev proces odlučivanja
- Reinforcment learning
- Pacman igrica
  - UC Berkeley, Into to AI
  - http://ai.berkeley.edu/reinforcement.html
  - 7 manjih zadataka



Markovljev proces odlučivanja

- Lavirint
- Ne krećemo se uvek kako smo planirali (noisy movement)
- Dobijamo nagradu pri svakom potezu
  - Mala nagrada za preživljavanje
  - Velike nagrade dolaze na kraju
- Cilj – maksimalna suma nagrada
- MDP definišemo
  - Skupom stana S
  - Startnim stanjem
  - Skupom akcija A
  - Prelascima T(s,a,s') ili P(s' | s,a)
  - Nagradama R(s,a,s') sa umanjenjem gama





Vrednost stanja

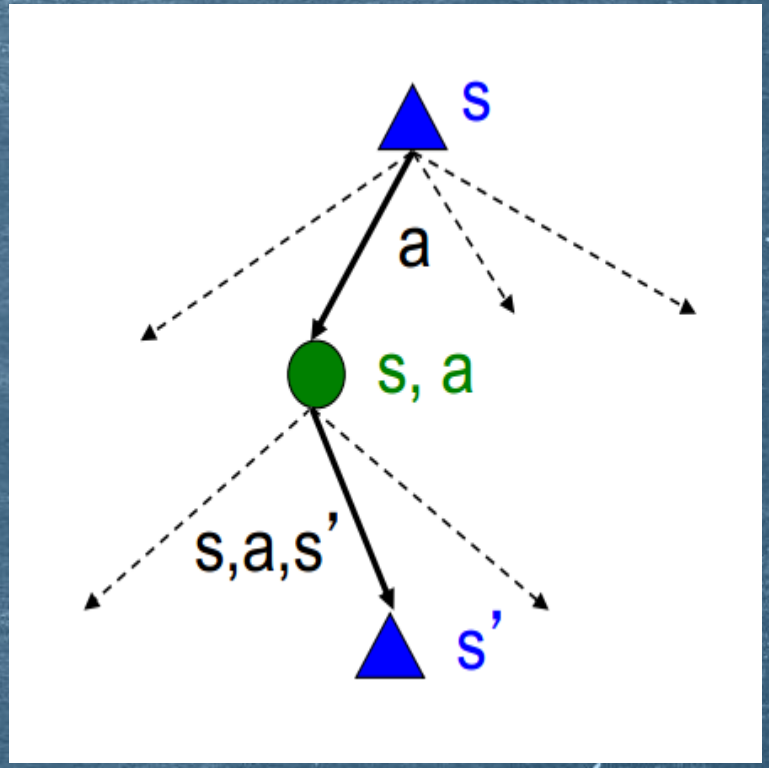
$$V^*(s) = \max_a Q^*(s,a)$$

$$Q^*(s,a) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

Iteracija vrednosti:

$$V^*(s)_{k+1} = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*{}_k(s') \right]$$

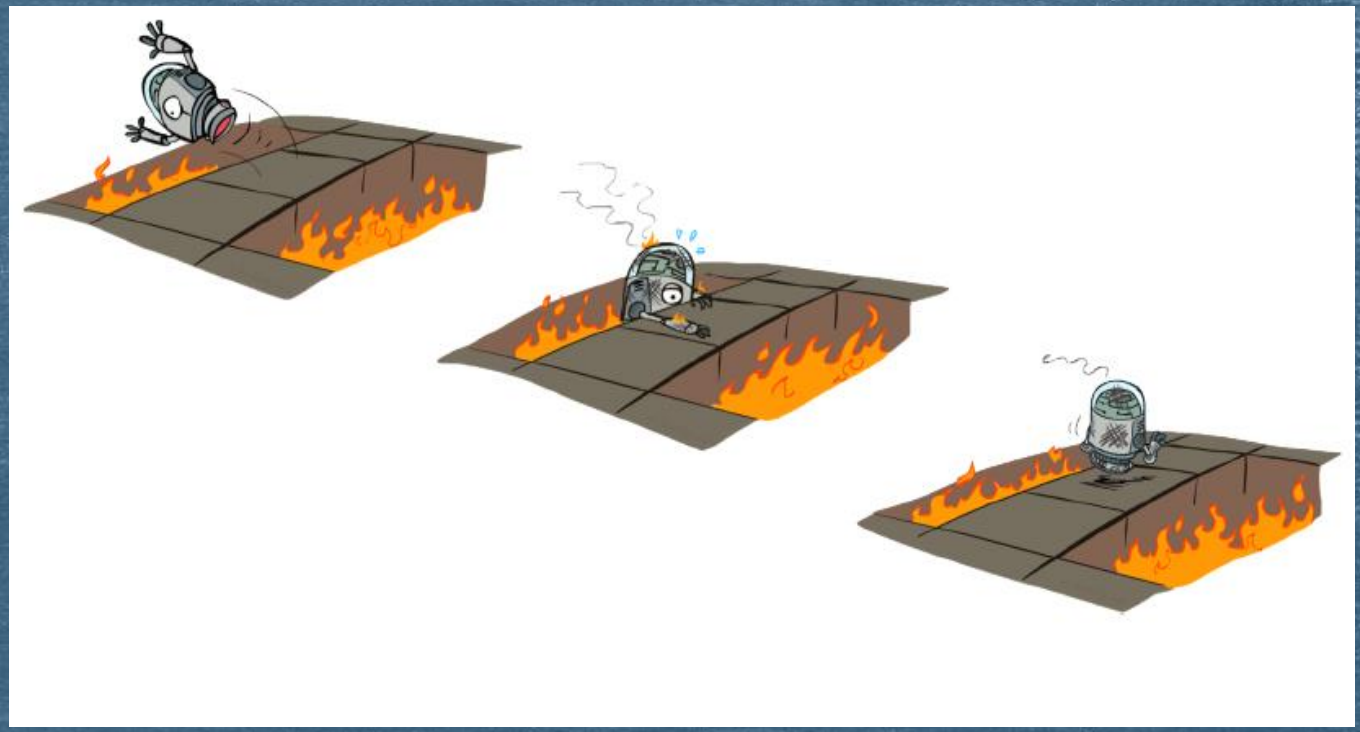


Pacman zadaci

- Zadatak 1
- Zadatak 2
- Zadatak 3
- http://ai.berkeley.edu/reinforcement.html

Aktivno pojačano učenje

- Ne znamo R(s,a,s'), T(s,a,s')
- Sami donosimo odluke
- Exploration vs. exploataion



QLearining

Iteracija Q-vrednosti zasnovana na uzorku:

$$Q_{k+1}(s,a) \leftarrow \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q_k(s',a') \right]$$

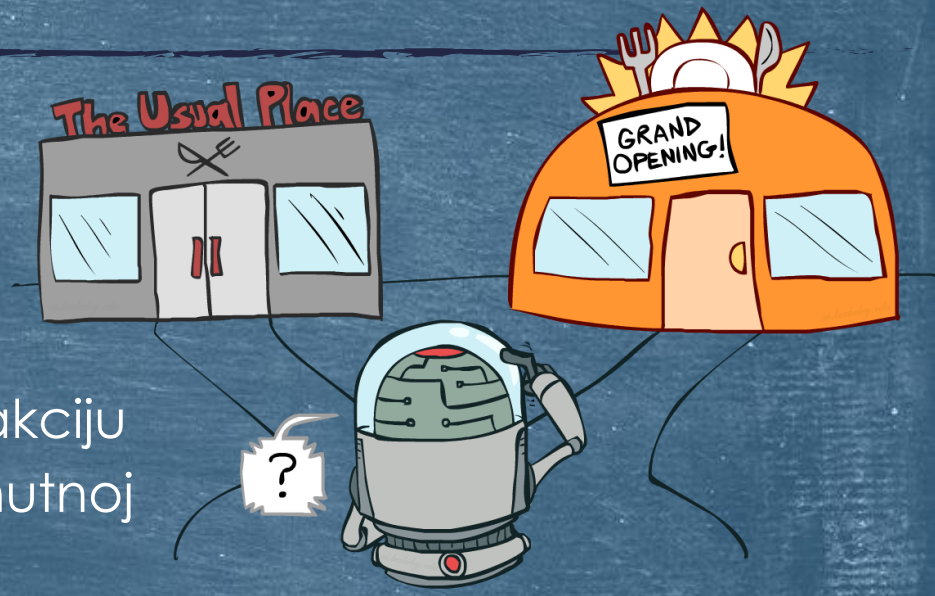
Za uzorak (s,a,s',r) računamo novu vrednost stanja:

$$sample = R(s,a,s') + \gamma \max_{a'} Q(s',a')$$

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha)\,[sample]$$

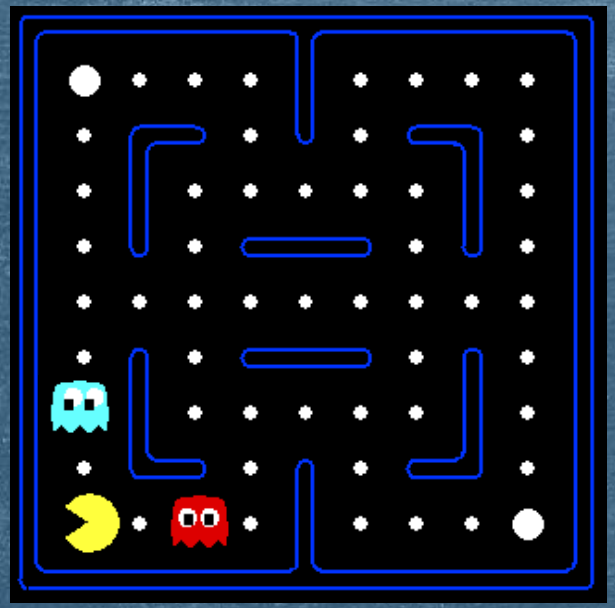
Istraživanje ili eksploatacija

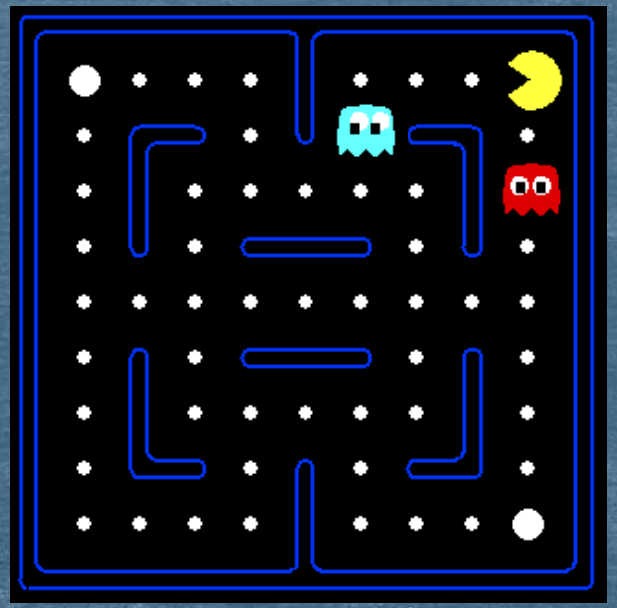
- Kako istraživati?
  - Svaki put pacamo novčić
  - Sa malom verovatnoćom ϵ biramo slučajnu akciju
  - Sa verovatnoćom $1 - \epsilon$, biramo akciju po trenutnoj politici
  - Vremenom smanjivati ϵ

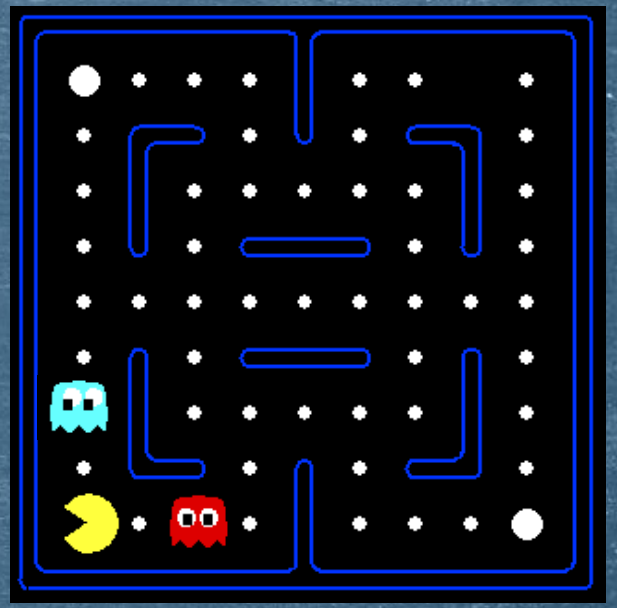


Aproksimacija Q-vrednosti

- Problem sa Q-vrednostima







Reprezentacija svojstvima (feature-based)

$$Q(s,a) = w_1 f_1(s,a) + w_2 f_2(s,a) + \ldots + w_n f_n(s,a)$$

$$\text{difference} = \left[r + \gamma \max_{a'} Q(s',a') \right] - Q(s,a)$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha \, [\text{difference}]$$

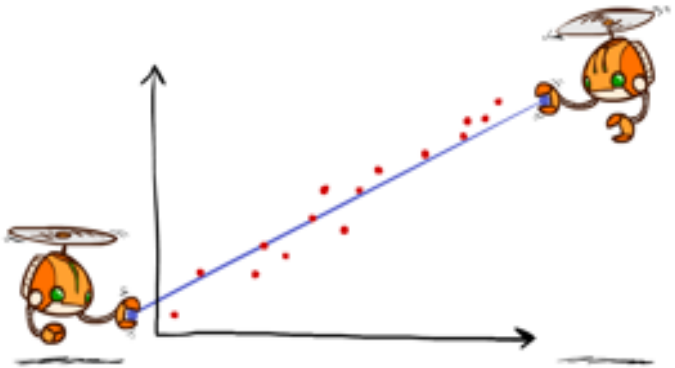
$$w_i \leftarrow w_i + \alpha \, [\text{difference}] \, f_i(s,a)$$

Metoda najmanjih kvadrata

$$\text{error}(w) = \frac{1}{2}\left(y - \sum_k w_k f_k(x)\right)^2$$

$$\frac{\partial\ \text{error}(w)}{\partial w_m} = -\left(y - \sum_k w_k f_k(x)\right) f_m(x)$$

$$w_m \leftarrow w_m + \alpha\left(y - \sum_k w_k f_k(x)\right) f_m(x)$$



$$w_m \leftarrow w_m + \alpha\left[r + \gamma \max_a Q(s', a') - Q(s, a)\right] f_m(s, a)$$